REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
21-07-2014	Conference Proceeding	-

4. TITLE AND SUBTITLE
Modeling Attrition in Organizations from Email Communication

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W911NF-11-C-0216

5c. PROGRAM ELEMENT NUMBER
1M30BM

6. AUTHORS
Akshay Patil, Juan Liu, Jianqiang Shen, Oliver Brdiczka, Jie Gao, John Hanley

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Palo Alto Research Center (PARC)
3333 Coyote Hill Road

Palo Alto, CA 94304 -1314

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60135-NS-DRP.10

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Modeling people's online behavior in relation to their real-world social context is an interesting and important research problem. In this paper, we present our preliminary study of attrition behavior in real-world organizations based on two online datasets: a dataset from a small startup (40+ users) and a dataset from one large US company (3600+ users). The small startup dataset is collected using our privacy-preserving data logging tool, which removes personal identifiable information from content data and extracts only aggregated statistics such as word frequency counts and sentiment features. The privacy preserving measures have enabled us to recruit participants to support

15. SUBJECT TERMS
social network,churn prediction,organization,email

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Oliver Brdiczka
UU	UU	UU	UU		19b. TELEPHONE NUMBER 650-812-4425

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Modeling Attrition in Organizations from Email Communication

ABSTRACT

Modeling people's online behavior in relation to their real-world social context is an interesting and important research problem. In this paper, we present our preliminary study of attrition behavior in real-world organizations based on two online datasets: a dataset from a small startup (40+ users) and a dataset from one large US company (3600+ users). The small startup dataset is collected using our privacy-preserving data logging tool, which removes personal identifiable information from content data and extracts only aggregated statistics such as word frequency counts and sentiment features. The privacy-preserving measures have enabled us to recruit participants to support this study. Correlation analysis over the startup dataset has shown that statistically there is often a change point in people's online behavior, and data exhibits weak trends that may be manifestation of real-world attrition. Same findings are also verified in the large company dataset. Furthermore, we have trained a classifier to predict real-world attrition with a moderate accuracy of 60-65% on the large company dataset. Given the incompleteness and noisy nature of data, the accuracy is encouraging.

Conference Name: 2013 International Conference on Social Computing SOCIALCOM

Conference Date: September 08, 2013

Modeling Attrition in Organizations From Email Communication

Akshay Patil* Juan Liu† Jianqiang Shen† Oliver Brdiczka† Jie Gao* John Hanley†

* Department of Computer Science, Stony Brook University, Stony Brook, NY. {akshay, jgao}@cs.stonybrook.edu
† Palo Alto Research Center, Palo Alto, CA. {Juan.Liu, Jianqiang.Shen, Oliver.Brdiczka, John.Hanley}@parc.com

***Abstract*—Modeling people's online behavior in relation to their real-world social context is an interesting and important research problem. In this paper, we present our preliminary study of attrition behavior in real-world organizations based on two online datasets: a dataset from a small startup ($40+$ users) and a dataset from one large US company ($3600+$ users). The small startup dataset is collected using our privacy-preserving data logging tool, which removes personal identifiable information from content data and extracts only aggregated statistics such as word frequency counts and sentiment features. The privacy-preserving measures have enabled us to recruit participants to support this study. Correlation analysis over the startup dataset has shown that statistically there is often a change point in people's online behavior, and data exhibits weak trends that may be manifestation of real-world attrition. Same findings are also verified in the large company dataset. Furthermore, we have trained a classifier to predict real-world attrition with a moderate accuracy of 60-65% on the large company dataset. Given the incompleteness and noisy nature of data, the accuracy is encouraging.**

I. Introduction

As computer and communication technologies become integrated into people's daily life, many actions and decisions that used to happen in real-world are now carried out on online platforms. For instance, email has become a ubiquitous means of communication, and social groups are increasingly important for people to share information with their real-world friends. The same happens in corporate environment. An increasing number of companies now rely on technologies such as Facebook-like social platforms, twitter-like information sharing platforms, and instant messages for their daily operations. As real-world actions manifest into online data, modeling online behavior in relation to the real-world behavior becomes an interesting and important research topic in both computer science and sociology. Researchers try to address questions such as how people's real-world social context shows up in online social data, whether and how one's action/decision is influenced by online interactions, and whether such manifestation and influence can be modeled and predicted. Many approaches have been proposed and experimented with, for instance, network analysis of social network graphs, content analysis of email text, and temporal analysis of email traffic. The topic of this paper is along the same line. We study online social interactions, in particular, email communications, and use them to understand real world behaviors of the involved individuals.

In this paper, we focus on the problem of understanding attrition within real-world organizations. Attrition rate, also called churn rate, in its broadest sense, is a measure of the number of individuals or items moving out of a collective over a specific period of time [1]. The term is used in many contexts, for example, referring to users switching to different service providers under a subscriber-based service model, players changing affiliation in online multiplayer games, user involvement and movement in online social networking platforms, and employee turnover within an organization. Attrition rate reflects an important aspect of group stability and thus is highly related to the profitability of a business. Analyzing, modeling, and predicting churn is of great practical significance and has also been extensively studied in research communities. Modeling and predicting attrition in a corporate setting is of particular practical importance.

However, the problem of modeling churn in an organization has received very little or nearly no attention in terms of concrete data analysis. This is largely due to the difficulty in data collection. Privacy concerns are often a major hurdle to data collection and need to be properly addressed. In our work, we have spent extra effort to guard personal information and designed a two-stage approach: (1) careful anonymization in the preprocessing stage to remove personally identifiable information (PII) such as name, address, email address and all numbers, and (2) the innovative design of an email logger, which processes emails and retains only aggregated statistics. Furthermore we assure that the aggregated statistics are sufficiently abstract such that the original content and/or meaning cannot be reconstructed from the feature set. These measures mitigate the privacy concerns and are key to the success of our data collection.

Besides the challenges of data collection, modeling and predicting attrition in a corporate environmetn is difficult from a data analysis perspective. First, any collected online data is inherently incomplete. It is impossible to collect a complete dataset of an individual in his/her social context – whom the individual is in contact with, what he/she is up to, the person's emotional state, etc. Any data collection approach is a best-effort approach. To make things worse, the ground truth, i.e., whether and why an employee is leaving the company, is noisy in nature or sometimes even lacking. There are various possible reasons attributing to an employee's departure, for instance, a voluntary leave (e.g., better opportunities offered by a competing company, taking time off for personal reasons, etc) or an involuntary one (management administering a lay-off as an organizational cost-cutting measure, or being fired for job disfunction). The exact underlying reason is unobservable. This makes the problem of attrition modeling and prediction difficult. However, we argue that, despite the diversity in

978-0-7695-5137-1/13 $26.00 © 2013 IEEE
DOI 10.1109/SocialCom.2013.52

possible reasons, almost all of them likely stem from job dissatisfaction (job mismatch, feeling devalued, stress, lack of support, and so on). This has been validated by extensively studied scenarios in social science literature [2], [3].

In this paper, we present a data-driven approach to attrition modeling and prediction. To the best of our knowledge, this is the first study of its kind to tackle this problem using real-world data. We obtain real workplace activity and communication data from two sources, a small startup company (43 employees) and a large US company ($3600+$ employees), as a platform for studying this problem. From the dataset, we extract a rich feature set and first perform correlation analysis to discover which features are strong in correlation with the departure. Some findings are expected and well in alignment with qualitative findings in social science studies. For instance, quitters may be involved more in external communication (i.e., with others outside of the company) and less in internal communications (with colleagues within the company). Quitters may initiate fewer email conversations and instead forward more onto colleagues. Our analysis has also discovered some unanticipated findings. For instance, we originally anticipated that through emails people may convey less positive sentiment and more negative sentiment as they disengage from the company, but the correlation analysis indicates that both negative and positive sentiment go down, and people become generally less expressive. This phenomenon is observed in both datasets. From these correlation observations, we build a model to predict if and when an employee is likely to quit the company. Our predictor achieves a modest accuracy of roughly $60 - 65\%$ prediction accuracy over the dataset from the large company. Giving the noisy nature of the data and the difficulty in attrition prediction, the prediction accuracy is encouraging.

The main contribution of our paper is as follows:

1) We have designed a privacy preserving data logging and feature extraction approach to capture a rich online behavior dataset free of personally identifiable information.
2) Through correlation analysis, our study shows that people's online behavior often exhibits a change point due to quitting intent. Though such change point has been speculated in various social studies, our work is the first to validate its existence from online data traces.
3) Over a large dataset in a corporate environment, we have built a predictive model predicting incipient quitting behavior based on online data with a moderate prediction accuracy.

II. Related Work

Here we briefly summarize related work on attrition from social science studies, business surveys, and data analysis.

A. Qualitative Findings from Social Science Studies

Social science studies have discovered that company size, industry and pay scales play a key role in determining attrition rate [1]:

- Larger companies tend to have lower rates of attrition.
- Industries that employ a large number of unskilled labor have a higher rate of attrition as compared to the ones that largely require skilled labor.
- Attrition rate is the highest amongst lowest paying jobs and vice-versa.

These findings have provided us with qualitative insights, but have not yet reached the level of mathematical precision that can be deployed to perform churn modeling and prediction. Our paper aims to fill this gap.

B. Studies on Churn Rates

In the scenario of subscriber based services, subscribers may leave a service provider for a number of reasons, including customer dissatisfaction, cheaper/better services/products provided by the competitors, or better marketing of products by competitors. Since this scenario has a direct impact on the profitability of a company, often companies come up with strategies to stem this exodus of subscribers. Strategies such as, creating barriers for subscribers to prevent them from switching or providing incentives such as loyalty programs are utilized for this purpose. Companies go this length because the cost of retaining an existing customer is far less than acquiring a new one [4]. The telecommunications industry gives special attention to this problem [5], [6], [7], [8], [9], [10], [11], [12]. This is due to the low barriers involved in switching service providers. The problem has been studied in other service-based industries as well, such as banking [13], ISPs [14], credit cards [15], insurance [16] and P2P networks [17].

A particular example is the video-gaming industry, a huge and growing market around the world, valued at about $65 billion in 2011 [18]. In order to generate revenues, it is critical that a given game is able to (a) attract new gamers and (b) retain gamers that are already playing the game. Coelln [19] talks about different metrics that can be used to gauge the success of a game in retaining current gamers. The strategies used by various gaming providers in retaining loyal users is discussed in [20]. Various studies [21], [22] have also analyzed the problem of churn prediction in online games.

The same problem of reducing churn rates appears in the scenario of social networking platforms. In recent years there has been a proliferation in the kind and number of social networking platforms available to people for interacting in the online space. For a social networking platform to remain active, it is important that it is able to attract users to contribute over a long period of time. Thus identifying people who are likely to churn early would allow the platform to investigate such users and make changes (or add features) to rectify such a situation. In [23], [24], the authors try to identify user features that can lead to churn in social networks. They then use these features to train classifiers for churn prediction achieving fair accuracy.

Most of the work listed above considered customers leaving/quitting the services as separated, independent events. They seek reasons from the internal of an individual. The proposed method for lowering churn rate is either by improving user satisfactory and providing better incentives, or by raising

the bar of switching services and imposing penalties. Nevertheless, in reality customers do not make quitting decisions independently. There is a hidden, yet powerful social factor behind user decision-making processes. It is often the case that a customer decides to subscribe/unsubscribe a service based on the decisions of his/her social friends. The aspect of social relationships in modeling and predicting churns is only studied until very recently. In our previous work [25], we looked at a related problem of predicting departures from groups and whether such departures are likely to cause damage to a social group. We performed the analysis using World of Warcraft (WoW), the most popular online role-playing game, as a platform. The analyses from real-world data sets indeed demonstrate a clear social influence when people make decisions on joining/quitting social groups.

C. Career Switch Modeling

A recent paper by Wang *et al.* [26] proposes a probabilistic model for career switching, to help a career recommender system in providing recommendations at the right time, i.e., to identify the time-period when a user is likely to be susceptible to making a career-switch. It models the duration between two successive job-related actions (such as a promotion or a churn event) using a proportional hazard model [27]. Proportional hazard is a technique that originates from reliability theory, describing the life span of a component (in this case the length that the user remains at the same job level) as a function of a baseline duration modulated by exponentials of a set of related factors. The paper then fits the proportional hazard model to a job application database from LinkedIn[1].

Our work is fundamentally different from [26]. For instance [26] models using a survival model on tenure, i.e., for how long is an individual expected to remain in the same position, while our approach examines data traces and discovers which data features may point towards a manifestation of an underlying attrition. Unlike [26] that uses tenure to decide when to pop job recommendation to users, our method predicts incipient departure based on features that may capture precursors of one's departure.

III. Analysis on Startup Email Dataset

A. Dataset

For attrition modeling, we recruited 43 participants from a research lab to share their social interaction data. The participants have diverse job roles: mangers, individual contributors, and administration staff. About half of the participants were associated with an internal spin-off startup company, which had not been successful in its business venture and later got incorporated back into the research lab. During the business turmoil, a significant portion of the employees left the company. From a supervised learning perspective this is an ideal dataset for analyzing quitting behavior, with positive and negative data samples and ground truth of who quit the company and when.

[1]http://www.linkedin.com

As we mentioned earlier in Section I, one major barrier to attrition analysis is data collection. We have designed an email feature extraction tool that respect users' privacy. Our tool can be deployed as a software agent installed on participant's PC to extract features from Outlook emails. Our tool takes two steps to protect privacy: (1) Participants are assigned a random ID, and only this ID is used to identify the participant, and all PII such as name, address, email address, and all numbers are ignored from email content. (2) Email content is processed to extract aggregated features such as word frequency counts. No raw content is logged in a feature set, and hence the original content cannot be reconstructed from the feature set. Furthermore, upon completion of feature extraction, the software agent uploads only aggregated features such as word frequencies onto an encrypted server and uninstalls itself.

For PII removal, emails go through a set of carefully designed pre-processing steps. First, reply lines and signature blocks are detected using regular expression and conditional random field techniques, as describe in [28]. The identification of reply lines and signature blocks serves two purposes. First, the reply lines facilitate email thread reconstruction, e.g., A sends an email to B, which B then replies to A and cc to C. By capturing email threads we can formulate an email graph and analyze its structure. Secondly, reply lines and signature blocks often contain personal information and sometimes redundant or irrelevant content. For instance, reply lines often quote original text from a previous message, and signature blocks sometimes have quotes from famous people. We remove these content portions when extracting content features.

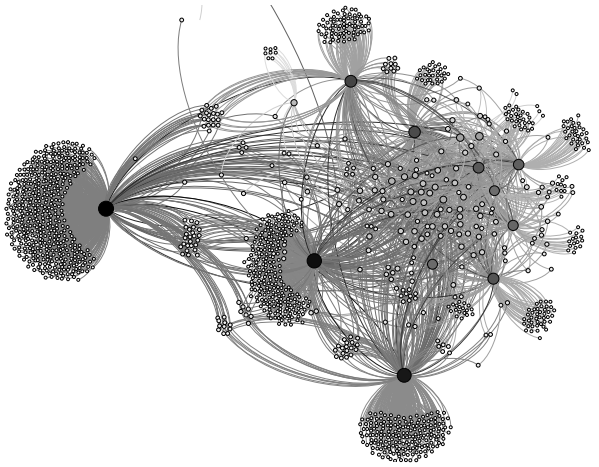

Fig. 1. Startup Email Graph. The larger, darker dots indicate internal employees, whereas the smaller, lighter dots indicate external email addresses.

B. Feature Set

Loosely speaking, we extract three categories of features: (1) meta-features, (2) graph structure features, and (3) content features.

Meta-features summarize various aspects of email usage: when an email send/receive event occurs, how many emails one sends a day, how many friends (defined as distinct people with email correspondence) one may have, how many internal

(to the company) emails/friends vs. external, and how many email correspondences are done outside of regular work hours.

Graph structure features reflect topological characteristics of email communication. All participants' email archives are used collectively to generate a dynamic email graph, where nodes are the participants, and edges are emails. Figure 1 visualizes the email graph that emerges from the communication traces. It is of interest to analyze the structure features of any given node, such as its degree or weighted degree, how tightly a node's local neighborhood is connected, and how balanced or skewed one's communication is within its neighborhood.

Content features are computed using text analysis techniques described in [29]. Here we briefly summarize the features:

Word statistics: This includes bag-of-words frequency counts for common words and the frequency counts for part-of-speech (POS) tags (e.g., noun, verb and adjective, superlatives, etc).

Sentiment features: A word can be positive, negative, neutral, or both positive and negative. The frequency for each category is counted. Negations are handled with special care.

Writing style: Professional emails often have a conventional structured format, with greeting in the beginning and closings at the end. Personal and professional emails often use smileys to express emotions. We add these to the feature set to investigate the expressiveness.

Speech act scores: One important use of work-related emails is to request or prompt certain actions, such as negotiation or task delegation. We use a pre-trained speech predictor [30] to predict six Speech Act scores: request, deliver, commit, propose, meet, and communicate data. Although not extremely accurate (around 70% F1 scores), the Speech Act scores make a good feature set due to its semantic importance, especially in work-related content.

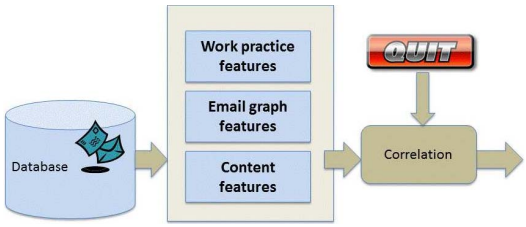

Fig. 2. Quitting dynamics analysis from email features for the Startup Dataset.

Figure 2 outlines the methodology for our analysis. From the dataset, meta-features, graph structure features, and content features are computed and collated into a feature set. They are then correlated with quitting ground truth. Due to the limited population size (43 participants total), we cannot train a reliable predictive model, but nevertheless it is important to address questions such as whether and how a quitting decision manifests into observation features, and which features are informative and indicative of quitting.

C. Class Labels

For investigating quitting dynamics, we first associate labels to differentiate multiple stages of one's employment. Given a participant, we take his/her employment period and segment it into 4 segments:

A warm-up period (class label 0): The first 6 weeks of employment is labeled as the warming up period. Email events in this period are discarded because they are not representative of one's regular work. Rather, the participant is communicating to familiarize with his/her environment.

An exit period (class label 3): The last 3 weeks of employment is labeled as the exit period. The employee has already made a decision to quit the company. The activities during this phase are usually wrapping up existing work and hand-off to colleagues.

A first & second working period (class labels 1 & 2): Cutting out the warm-up and exit periods, the rest of one's employment is divided into two equal-length segments. It is of interest to see whether one's email behavior changes during these two working periods (transition $1 \rightarrow 2$) and as he/she starts to exit (transition $2 \rightarrow 3$).

D. Correlation Analysis

Table I reports email features with non-negligible correlation coefficient with the employment stage labels. The second column ("transition $1 \rightarrow 2$") lists the correlation coefficient as participant transits from his/her first employment period to the second. The third column ("transition $2 \rightarrow 3$") lists the correlation coefficient as participant transits from the second employment period to the exit period. A positive correlation coefficient indicates that a general increase of feature value as participant progresses through the employment periods, while a negative correlation indicates a generally decreasing feature value.

The correlation coefficient for the meta-features exhibits a very distinct change pattern. For instance, as participants transit from stage 1 to 2, the number of emails generally goes up, but as participants transit from stage 2 to 3, the number of emails drops. Same pattern persists for a number of other features. In general people's internal email activities (internal with respect to the company) tend to drop, with fewer emails and fewer recipients. Also fewer activities after the normal working hours are observed. Fewer emails with attachments are observed. All these indicate generally less job-related activities. Furthermore, the number of forward emails drops but to a much lesser extent than the number of all emails. Percentage-wise the quitting participant is doing more email forwards and less originals or replies. This is also anticipated as people disengaging themselves from work. The email forwards are probably delegation and handoff to colleagues.

Among the feature set for graph structure, the entropy feature is computed over the empirical distribution of email frequency across recipients. It measures how evenly or skewed the participant communicates with email recipients. It takes a high value if the participant communicates with friends evenly, and a low value if his/her email activities are confined to a small subset of recipients. Table I shows that entropy drops, indicating that the communication is more skewed to a

TABLE I

CORRELATION BETWEEN EMAIL FEATURES AND GROUND TRUTH. THE COLUMN "TRANSITION $1 \rightarrow 2$" LISTS THE CORRELATION COEFFICIENT AS PARTICIPANT TRANSITS FROM FIRST EMPLOYMENT PERIOD TO THE SECOND. THE COLUMN "TRANSITION $2 \rightarrow 3$" LISTS THE CORRELATION COEFFICIENT AS PARTICIPANT TRANSITS FROM THE SECOND EMPLOYMENT PERIOD TO THE EXIT PERIOD. CORRELATION COEFFICIENTS WITH SIGNIFICANT AMPLITUDE (> 0.05) ARE HIGHLIGHTED AS FOLLOWS, RED & ITALIC FOR NEGATIVE CORRELATION, AND BLUE & BOLD FOR POSITIVE CORRELATION.

Class	Feature	Transition $1 \rightarrow 2$	Transition $2 \rightarrow 3$	Change Point
Meta-features	number of emails	**0.279**	*-0.252*	Fewer emails and internal communications
	number of distinct recipients	**0.318**	*-0.199*	
	number of emails internally sent	**0.279**	*-0.252*	
	number of internal recipients	**0.391**	*-0.199*	
	percentage of internal recipients	**0.109**	*-0.074*	
	number of after hours emails	**0.118**	-0.040	Change in work habits, fewer after-hours activity
	number of attachments	0.031	*-0.101*	
	average number of TO recipients	**0.126**	*-0.069*	Fewer composed or replied emails, increased percentage of forwarded emails
	average number of CC recipients	**0.318**	*-0.065*	
	number of forwarded emails	**0.155**	-0.017	
Graph Structure	entropy over email recipients	**0.284**	*-0.146*	More skewed communication, fewer cliques
	clustering coefficient	**0.127**	*-0.228*	
	weighted degree	**0.333**	*-0.108*	
Content	number of exclaimation marks	**0.117**	*-0.054*	Reduced expressiveness in communication
	number of emails with positive polarity	**0.257**	*-0.244*	
	number of emails with negative polarity	**0.279**	*-0.225*	
	number of emails with rare/complex words	**0.253**	*-0.258*	
	number of question marks	0.027	*-0.085*	No apparent changes
	number of POS Tags JJS (Superlatives)	*-0.061*	-0.046	

smaller subset. This is not surprising as we anticipate quitters to communicate more with close friends and less to the broader set of colleagues. Clustering coefficient [31] is defined as the ratio of triangle cliques over the number of possible triplets. It takes value 1 if the participant's local neighborhood is fully connected and 0 if the participants live in a star topology. Table I shows that the clustering coefficient generally drops. This is probably because the participant is involved in more external and less internal communication. The internal communication community is expected to be tightly knit, hence closer to a fully connected graph, while the external network is only connected through the participant and hence exhibits more of a star topology.

Analysis over content features turned out to be somewhat surprising. We had originally anticipated that positive sentiment may go down and negative sentiment may go up as people plans to exit. Data seems to suggest that both positive and negative sentiments go down. In addition, exclamation marks and rare/complex words are used less often than before. In general people become less expressive and distant themselves from controversial expressions. On the other hand there is no significant change observed in terms of questioning and the frequency of using superlatives. Although a change point has long been speculated and empirically observed in social science studies, our work is the first to report its existence in real-world email data. These preliminary results encourage us to look into the possibility of constructing predictive models to further analyze quitting behavior in larger datasets.

IV. ANALYSIS ON LARGE COMPANY DATASET

A. Dataset

The second dataset comes from a large US company with several tens of thousands of employees (hereafter we will refer to this dataset as the "Large Company Dataset"). The company has adopted a logging mechanism for their employees, logging activities on company operated PCs. Some of the types of activities logged are email communications, logon activities, file-access activities, websites visited from work PC etc. For our problem, email communications serve as the most important source of information as compared to other activity data.

TABLE II

DATA STATISTICS FOR LARGE COMPANY DATASET

Statistic for Large Company Dataset	
Number of Target Internal Employees	3,615
Number of Other Internal Employees	23,672
Number of External People	86,240
Number of Email Communications	37,619,622
Number of Quitters Identified	566 (15.65%)
Time Range	14-May-2012 to 01-Oct-2012

While studying our problem we limit our attention to communication data for a large subset ($3600+$) of employees during a 20-week time period. Table II details the scale of the dataset. For analysis, we distinguish individuals into three categories. The first category includes individuals that belong to our targeted subset of employees, i.e., the ≈ 3600 employees in our dataset. These are the samples for our analysis. The second category are the ones that are employees of the company but not in our subject dataset. They are excluded for purposes of churn analysis. The last category are people external to the company. All communication between the first two categories of people are treated as "internal" communication and all other communication is treated as "external" communication. For the purposes of this analysis we have focused on email communication trace data without using actual email content for analysis. We exclude actual

email content from our analysis due to privacy concerns. Since the dataset does not have actual information on employees quitting the company (i.e. date & reason of departure), we will use a simple but effective heuristic to determine the approximate date of departure. We will also model the email communication data as a social network that can then be used to extract certain structural features that could be useful for churn prediction. We construct the social network by placing an edge between two people that have participated in an email communication.

B. Identifying Quitters

The dataset does not have the ground truth with respect to employees quitting the company (that is treated as proprietary information). Since we require ground truth in order to train and validate our approach, we need to be able to deduce the fact that an employee has quit the company from the given communication and activity data.

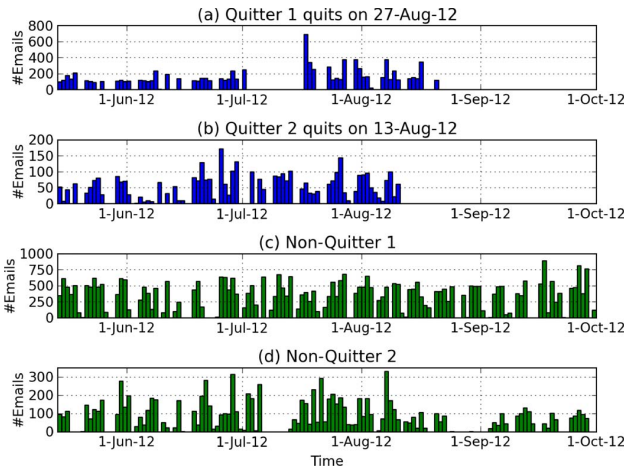

Fig. 3. Email Communication Plot for 4 employees. Top-2 employees (blue bars) are those that have been deemed to have quit the company based on the "Sent Email" heuristic". The 2-day gaps indicate lack of email activity on weekends. One can also see a temporary week-long absence for couple of the employees indicating vacation, sick leave etc.

We use a heuristic based on "Sent Emails". For a given employee we keep track of his/her sent emails at regular intervals of time (daily in our experiments). We can then classify an employee as having quit the company if we fail to see any emails being sent by this employee for a prolonged period of time such that this employee never sends out any further emails. The "Sent Email" heuristic should allow us to get reasonable ground truth regarding quitters. In our experiments we will say that an employee E has quit the company at time T if he has sent his last email at time T, followed by complete & permanent absence for at least 21 days. We set the minimum time duration for lack of sent emails as 21 days; that should take care of scenarios where employees are temporarily absent due to vacation, sick leave etc. While we realize that no heuristic would be perfect, we believe that our heuristic is effective in identifying most if not all quitters. We also believe that we can deduce the approximate date of departure for such employees by using the "Sent Email" heuristic. A drawback of our approach would be in cases where employees use (or switch to) machines that have not been registered for logging email and other activities. In such cases we might identify some employees as quitters even though they might still be employees. We consider such cases as noise; another reason that makes this problem harder than in other scenarios. Figure 3 shows the sent email activity plot for 4 employees, top 2 employees are deemed to have quit the company based on the "Sent Email" heuristic. In all, we identify 566 employees that have quit the company in the 20-week time period in the Large Company Dataset.

TABLE III
SPEARMAN'S RANK CORRELATION COEFFICIENT BETWEEN CLASS LABELS AND FEATURE VALUES. CORRELATION COEFFICIENTS WITH SIGNIFICANT AMPLITUDE (> 0.05) ARE HIGHLIGHTED AS FOLLOWS, RED & ITALIC FOR NEGATIVE CORRELATION, AND BLUE & BOLD FOR POSITIVE CORRELATION.

	Feature	Correlation Coefficient
Email	percentage of emails sent	0.0306
	percentage of distinct recipients	0.0294
	percentage of external emails	0.0263
	percentage of after hours external emails	0.0197
	percentage of after hours internal emails	-0.0197
	percentage of internal emails	-0.0263
	percentage of distinct senders	-0.0294
	percentage of emails received	-0.0306
	number of emails received	-0.0419
	number of distinct senders	-0.0426
	number of after hours internal emails	-0.0449
	number of after hours external emails	*-0.0541*
	number of after hours emails	*-0.0566*
	number of internal emails	*-0.0583*
	number of distinct recipients	*-0.0663*
	number of external emails	*-0.0667*
	number of emails sent	*-0.0670*
	number of emails	*-0.0703*
Social	percentage of external friends	0.0070
	percentage of internal friends	-0.0070
	number of friends emailed excessively	-0.0314
	number of emails per internal friend	-0.0490
	number of emails per external friend	*-0.0542*
	number of emails per friend	*-0.0569*
	number of internal friends	*-0.0804*
	number of external friends	*-0.0920*
	number of friends	*-0.0930*
Structural	clustering coefficient	-0.0453
	email entropy	*-0.0695*
	weighted degree	*-0.0706*

C. Feature Set

The first step towards churn analysis entails extracting a rich set of features from the communication dataset. Table III provides a list of features that we compute from the dataset. Most features are similar to those described in Section III, except that Meta-features are further split into two categories:

Email Features: The features are extracted from email communication data and form the bedrock of our analysis. Email

communications data provide the a large set of features that can be used to detect changes in workplace behavior.

Social Features: These are features that are a direct byproduct of constructing the email social graph. These features are useful in capturing statistics on one's ego network.

D. Feature Importance & Correlation

The observation dataset is organized into temporal snapshots, sampled at weekly intervals i.e. we compute the set of features for every employee at weekly intervals. Furthermore, we label each feature sample for an employee as being a "quitting" or "not quitting" sample. If an employee is to quit the company within the next 3 weeks, we classify the feature sample for that employee at that particular time as "quitting", if not then the sample is classified as "not quitting". Labeling in such a manner allows us to train classifiers for the churn prediction problem. Also, it allows the classifiers to uncover any precursors that lead to an employee departure. Overall there are about 50,000 feature samples, out of which about 1713 (3.42%) are samples with "quitting" label.

Table III reports the Spearman's Rank Correlation Coefficient between each feature and the class labels (1 for quitting, 0 for not quitting). We prefer Spearman's correlation coefficient over Pearson's correlation coefficient to mitigate issues that may arise due to outliers and skewed nature of the data. Some features exhibit distinct trends; following are the observations from the correlation analysis,

- An employee that is about to quit the company will participate in fewer emails (number of emails sent, number of emails received).
- He/she will communicate with a selected (fewer) group of people (number of friends, number of external friends, number of internal friends, email entropy, weigted degree, clustering coefficient).
- There is a marginal increase in his/her percentage of external emails i.e. emails sent to people outside the company.
- Overall such an employee is likely to be less communicative as compared to other employees.

Table IV shows the top-10 features as ranked by the information gain criterion. This indicates how informative a given feature is with respect to deducing the class label (quitting or not quitting). This is consistent with Table III regarding which features are most important for churn prediction.

TABLE IV
TOP TEN FEATURES IN DESCENDING ORDER OF INFORMATION GAIN.

Rank	Feature	Category
1	number of external friends	Social
2	number of friends	Social
3	number of emails sent	Email
4	number of distinct recipients	Email
5	number of emails	Email
6	number of internal friends	Social
7	weighted degree	Structural
8	number of external emails	Email
9	email entropy	Structural
10	number of emails per friend	Social

E. Predicting Quitters

Given the feature samples along with the class labels, we train classifiers for the churn prediction problem. Due to the skewed nature of the data (i.e. very few "quitting" samples), we first randomly select equal number of samples from both classes. We conduct experiments using a wide range of supervised learning techniques, with Bagging providing the best accuracy. We are able to achieve accuracy levels that range between 58-63% using various classifiers such as Naive Bayes, Decision Trees, Random Forests & Bagging. Figure 4 provides the precision, recall & f-measure numbers for each of the classifiers. Table V shows the detailed results obtained by using Bagging. One can see that we are able to achieve a modest accuracy of about 63% when predicting churn in an organization. The results also indicate that predicting churn in an organization seems to be a much harder problem than predicting churn in other well studied scenarios.

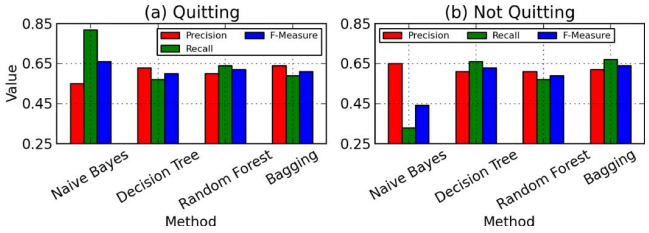

Fig. 4. Prediction Results using different classification techniques. The figures plot precision, recall & f-measure for the "Quitting" (left) & "Not-Quitting" (right) classes. The classification accuracy ranges between $58-63\%$ for the different techniques, with Bagging proving to be marginally better than the remaining methods.

TABLE V
PREDICTION ACCURACY USING BAGGING. BAGGING[32] IS AN ENSEMBLE METHOD WHICH IMPROVES THE CLASSIFICATION ACCURACY THROUGH SAMPLING AND MODEL AVERAGING.

Class	Accuracy	Precision	Recall	F-score
Quitting		0.641	0.588	0.613
Not Quitting	62.96%	0.62	0.672	0.645
Weighted Average		0.631	0.63	0.629

V. CONCLUSION

Through analysis on the startup email dataset and the large company dataset, we are able to make the first (albeit modest) headway into the problem of churn modeling prediction in an organization. Though the results are still preliminary, our analysis establishes a data-driven correlation model between people's online behavior and their real-world company churn events. We have identified informative features and found weak trends, indicating that employees become less communicative and less expressive before they are likely to quit the company. Our analysis also has verified the existence of a change point in people's work-related emailing behavior. Furthermore, through analysis on the large company dataset, we can use the weak trends to predict incipient departure with a moderate accuracy.

This effort is inscribed in a broader scope of research seeking to establish correspondence between people's real-world context and online social behavior. We hope that the method presented in this paper can be extended to other similar problems as well. We plan to look into other social dynamics, such as influence propagation, dynamics between work place collaborators, and roles people take in their online social groups. Through such effort we hope to achieve a better understanding of the social space around us, so as to make online platforms more helpful to real-world users.

ACKNOWLEDGEMENTS

This research is funded in part by DARPA/ADAMS program under contract W911NF-11-C-0216. Any opinions, findings, and conclusions or recommendations in this material are those of the authors and do not necessarily reflect the views of the government funding agencies. We would also like to acknowledge contributions from our colleagues Robert Price, and Hoda Eldardiry for helpful advice throughout the project, and Kevin Raminhos and William Jennings for help with preprocessing the data.

REFERENCES

[1] "Churn rate," http://en.wikipedia.org/wiki/Churn_rate.

[2] D. Fields, M. E. Dingman, P. M. Roman, and T. C. Blum, "Exploring predictors of alternative job changes," *Journal of Occupational and Organizational Psychology*, vol. 78, no. 1, pp. 63–82, 2005.

[3] K. Jiang, D. Liu, P. F. McKay, T. W. Lee, and T. R. Mitchell, "When and how is job embeddedness predictive of turnover? a meta-analytic investigation." 2012.

[4] C. Hart, J. Heskett, W. Sasser *et al.*, "The profitable art of service recovery," *Harvard business review*, vol. 68, no. 4, pp. 148–156, 1990.

[5] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi, "Social ties and their relevance to churn in mobile telecom networks," in *EDBT*, 2008, pp. 668–677.

[6] J. B. Ferreira, M. Vellasco, M. A. Pacheco, and C. H. Barbosa, "Data mining techniques on the evaluation of wireless churn," in *In ESANN*, 2004, pp. 483–488.

[7] B. Huang, B. Buckley, and T. M. Kechadi, "Multi-objective feature selection by using nsga-ii for customer churn prediction in telecommunications," *Expert Syst. Appl.*, vol. 37, no. 5, pp. 3638–3646, May 2010. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2009.10.027

[8] S.-Y. Hung, D. C. Yen, and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, no. 3, pp. 515 – 524, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417405002654

[9] B. Masand, P. Datta, D. Mani, and B. Li, "Champ: A prototype for automated cellular churn prediction," *Data Mining and Knowledge Discovery*, vol. 3, pp. 219–225, 1999, 10.1023/A:1009873905876. [Online]. Available: http://dx.doi.org/10.1023/A:1009873905876

[10] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," *Trans. Neur. Netw.*, vol. 11, no. 3, pp. 690–696, May 2000. [Online]. Available: http://dx.doi.org/10.1109/72.846740

[11] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12 547–12 553, Dec. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2009.05.032

[12] Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Customer churn prediction using improved one-class support vector machine," in *Proceedings of the First international conference on Advanced Data Mining and Applications*, ser. ADMA'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 300–306. [Online]. Available: http://dx.doi.org/10.1007/11527503_36

[13] K. Coussement and D. V. den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Applications*, vol. 34, no. 1, pp. 313 – 327, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417406002806

[14] B. Q. Huang, M.-T. Kechadi, and B. Buckley, "Customer churn prediction for broadband internet services," in *Proceedings of the 11th International Conference on Data Warehousing and Knowledge Discovery*, ser. DaWaK '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 229–243. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03730-6_19

[15] G. Nie, G. Wang, P. Zhang, Y. Tian, and Y. Shi, "Finding the hidden pattern of credit card holder's churn: A case of china," in *Proceedings of the 9th International Conference on Computational Science*, ser. ICCS 2009. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 561–569. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01973-9_63

[16] K. Morik and H. Köpcke, "Analysing customer churn in insurance data: a case study," in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, ser. PKDD '04. New York, NY, USA: Springer-Verlag New York, Inc., 2004, pp. 325–336. [Online]. Available: http://dl.acm.org/citation.cfm?id=1053072.1053103

[17] O. Herrera and T. Znati, "Modeling churn in p2p networks," in *Simulation Symposium, 2007. ANSS '07. 40th Annual*, march 2007, pp. 33 –40.

[18] T. R. Corporation, "Factbox: A look at the $65 billion video games industry," http://uk.reuters.com/article/2011/06/06/us-videogames-factbox-idUKTRE75552I20110606, 2011.

[19] E. von Coelln, "The sticky factor: Creating a benchmark for social gaming success," http://www.insidesocialgames.com/2009/10/27/the-sticky-factor-creating-a-benchmark-for-social-gaming-success/, 2009.

[20] ——, "How big social games maintain their sticky factors," http://www.insidesocialgames.com/2009/11/04/how-big-social-games-maintain-their-sticky-factors/, 2009.

[21] J. Kawale, A. Pal, and J. Srivastava, "Churn prediction in mmorpgs: A social influence based approach," in *CSE (4)*, 2009, pp. 423–428.

[22] Z. Borbora, K. Hsu, J. Srivastava, and D. Williams, "Churn prediction in mmorpgs using player motivation theories and ensemble approach," *Proceedings of SocialCom-11*, 2011.

[23] M. Karnstedt, M. Rowe, J. Chan, H. Alani, and C. Hayes, "The effect of user features on churn in social networks," in *Third ACM/ICA Web Science Conference*, 2011. [Online]. Available: http://www.websci11.org/fileadmin/websci/Papers/189_paper.pdf

[24] G. Dror, D. Pelleg, O. Rokhlenko, and I. Szpektor, "Churn prediction in new users of yahoo! answers," in *Proceedings of the 21st international conference companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 829–834. [Online]. Available: http://doi.acm.org/10.1145/2187980.2188207

[25] A. Patil, J. Liu, B. Price, H. Sharara, and O. Brdiczka, "Modeling destructive group dynamics in on-line gaming communities," in *Proceedings of the 6th International AAAI Conference on WebBlogs and Social Media (ICWSM)*. AAAI, 2012, pp. 290–297.

[26] J. Wang, Y. Zhang, C. Posse, and A. Bhasin, "Is it time for a career switch?" in *Proceedings of the 23nd International World Wide Web Conference*. ACM, 2013, pp. 1377–1387.

[27] D. R. Cox and D. Oakes, *Analysis of survival data*. Chapman & Hall/CRC, 1984, vol. 21.

[28] V. R. Carvalho and W. W. Cohen, "Learning to extract signature and reply lines from email," in *Conference on Email and Anti-Spam (CEAS-04)*, 2004.

[29] J. Shen, O. Brdiczka, and J. Liu, "Understanding email writers: Personality prediction from email messages," in *The 21st Conference on User Modeling, Adaptation and Personalization (UMAP)*, 2013.

[30] W. W. Cohen, V. R. Carvalho, and T. M. Mitchell, "Learning to classify email into speech acts," in *Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, 2004.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[32] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.